RED WINE QUALITY PREDICTION

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Parth Sharma (181439)

Under the supervision of

Dr.Aman Sharma

to

Department of Computer Science & Engineering and Information Technology
**Jaypee University of Information Technology Waknaghat, Solan-173234,
Himachal Pradesh**

Candidate's Declaration

I hereby declare that the work presented in this report entitled **" Red wine quality prediction"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2021 to May 2022 under the supervision of **Dr Aman Sharma**
The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Sharma.

Parth Sharma

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Aman Sharma

Dr Aman Sharma

Assistant Professor (SG)

Computer Science

Dated: 13/05/2022

Acknowledgement

I would like to convey my heartfelt gratitude to everyone who made it possible for us to put this project together. I would like to convey our heartfelt gratitude to Dr. Aman Sharma, my final year project adviser, whose persistent encouragement and helpful recommendations aided us in successfully completing my project, particularly throughout the drafting of this report.I would also like to express our gratitude to the Jaypee University of Information Technology team, who allowed us to use all of the necessary equipment and materials to accomplish my project "Red Wine Quality Prediction Using Machine Learning." Last but not least, I would like to express my gratitude to my other classmates who assisted me with various areas of our project and provided solutions to my challenges. I am grateful for the advice I received from other faculty members, particularly during my project presentation, which has helped me enhance my presentation skills. I appreciate their suggestions and remarks during the presentation.

Table of Contents

List of Figures

List of Tables

Abstract

As the industrial revolution took place,civilisations and humans are evolving at a faster pace than ever seen before.To catch up with the increased supply demands ,goods are being made artificially which are also increasing the profits.As a result the quality of food is deteriorating which has lead to the increased risk of severe health problems in human being.We now need a system to forecast the quality of the drink and meal we are consuming.In this research paper we have focused on red wine quality.The dataset contains important features such as alcohol,residual sugar,density.Different measures were performed to evaluate our proposed framework such as Precision , Sensitivity etc.Our Proposed framework attained an accuracy of 98.36% which outperformed previous literature work.

Chapter 1

1.1 Introduction

According to the OIV, global wine consumption in 2020 would be 234 million hectolitres (Mhl). In comparison to 2019, this is a 3 percent (7 Mhl) decrease. Consumption has decreased for the third year in a row. It's at its lowest point since 2002[20].The United States, France, and Italy are the top three wine-consuming countries. Portugal, Italy, and France are the three countries with the highest per capita wine consumption[20].A diet high in low-quality foods raises your risk of chronic diseases, whereas a diet high in high-quality foods protects you.[1].Certain types of cancer are even influenced by diet. According to the World Health Organisation, a nutritious diet reduces the risk of malignancies such as colon, breast, and kidney.[2].Every year, an estimated 600 million individuals – about one in every ten people in the globe – become unwell after eating contaminated food, and 420 000 die, resulting in the loss of 33 million healthy life years (DALYs)[3].Alcohol usage at the age of 15 predicted weekly alcohol consumption and alcohol intake exceeding the prescribed level four years later.

The increased alcohol intake of young teenagers was not a passing fad. It was a pattern that continued throughout young adulthood, putting the teenagers at a higher risk of becoming long-term, large-scale consumers.[4].At the age of 19, at least 80% drank alcohol monthly, and 24% of men and 11% of women used alcohol in excess of the prescribed national limits, i.e. 21 weekly units of alcohol for males and 14 for women. Use of alcoholic drinks at the age of 15 increased the likelihood of weekly alcohol consumption at the age of 19 [odds ratio (OR)-values ranging from 1.11 to 3.53]. Drunkenness among 15-year-old boys and spirit use among 15-year-old girls showed the best predictive connection with excessive consumption at age 19.[OR = 2.44, confidence interval (CI): 1.38–4.29, respectively, OR = 1.97, CI: 1.15–3.38][4].Excessive alcohol intake is associated to a number of undesirable outcomes, including being a risk factor for diseases and health effects, criminality, traffic accidents, and, in some cases, alcohol dependence. Each year, 2.8 million people die prematurely due to alcohol usage around the world[22].For hundreds of years, red wine has been a component of social, religious, and cultural gatherings.

Monasteries in the Middle Ages believed that their monks lived longer because they drank wine on a regular, moderate basis.According to a report published in 2018[23], Drinking red wine in moderation has positive ties with: Trusted Source, although there are no official guidelines around these advantages, drinking red wine in moderation has positive linkages with:cardiovascular disease,atherosclerosis,hypertension,certain types of cancer,type 2 diabetes,neurological disorders,metabolic syndrome.Red wine, which is created from crushed black grapes, is a good source of resveratrol, a natural antioxidant found in grape skin[24]. Antioxidants help the body fight oxidative stress. Many diseases, including cancer and heart disease, have been linked to oxidative stress. Fruits, nuts, and vegetables are just a few of the antioxidant-rich meals available. Whole grapes and berries have more resveratrol than red wine, and because of the health hazards associated with alcohol consumption, receiving antioxidants from food is likely to be healthier than drinking wine. To receive enough resveratrol, people may need to drink a lot of red wine, which may cause more harm than good. When it comes to alcoholic beverages, however, red wine may be more beneficial than others.Whole grapes and berries have more resveratrol than red wine, and because of the health hazards associated with alcohol consumption, receiving antioxidants from food is likely to be healthier than drinking wine. To receive enough resveratrol, people may need to drink a lot of red wine, which may cause more harm than good[25].

The outline of the rest of the paper is as follows: In the second section we have included the literature review in which we referred to various research works and explained the viability and performance of the different algorithms related to heart disease prediction. In Section 3 we have explained different machine learning algorithms. In Section 4, the proposed framework has been explained in details including model selection, parameter setting, Experimental setup & proposed methodology. In Section 5, performance metrics, comparison of proposed framework with existing Machine Learning (ML) models & with existing literature is explained. Results are shown with respect to existing models & literature in this section. Section 6 contains the conclusion and future scope.

1.2 Problem Statement

The significance of each feature for the wine quality prediction is not yet quantified. And in terms of performance, the current accuracy is about 67.25%. Thus, in this research, we considered two aspects of the problems mentioned above. The first one is the study of the importance of the features for the prediction of wine quality.Secondly, performance of the prediction model can be improved using ensemble learning with other ordinary classifiers

1.3 Objective

The following research question and hypothesis are formulated.

1. What wine features are important to get a promising result?

The researchers have used classification algorithms but for the classification task hyperparameter tuning and Ensembled techniques were never used. Hypothetically, the current prediction model that has been obtained by researchers will be improved by using the ensemble technique.

To address the research question the following objectives are formulated :
1.To balance the dataset.
2. To analyse the impact of the features.
3.To optimise the classification models through hyperparameter
Tuning and ensemble learning.
4. To model and evaluate the approaches.

Machines could use this skill to improve their interactions with people by offering more appropriate responses. Finally, this is frequently a multidisciplinary endeavor that requires efficient computing and machine learning. Another project's purpose is to learn how these various professions are related and how they will bring solutions to difficult challenges.

In this report, we start with analysing the info provided and supporting various labels provided. This concept comes from our lifestyle because we will always identify the people around us and support their important features which we notice in day to day life. it'd be a decent choice since this method really reduces the computational complexity and will also achieve good accuracy.

1.4 Methodology

Stacking is a type of ensemble learning where base learners predictions are done on the bases of meta - learner through union of various algorithms.In the proposed algorithm top performing models are selected from all the algorithms and combined together to give even higher accuracy .The data is taken from UCI machine learning repository and the data is split into 80% training and 20% testing after correcting class imbalance through SMOTEENN and skewness is corrected by using PowerTransformer library.

The following attributes skewness is corrected because the value was high that 0.75:chlorides, total sulfur dioxide, residual sugar,free sulfur dioxide, sulfates, volatile acidity.Figure 6 shows our proposed methodology as a flow diagram.

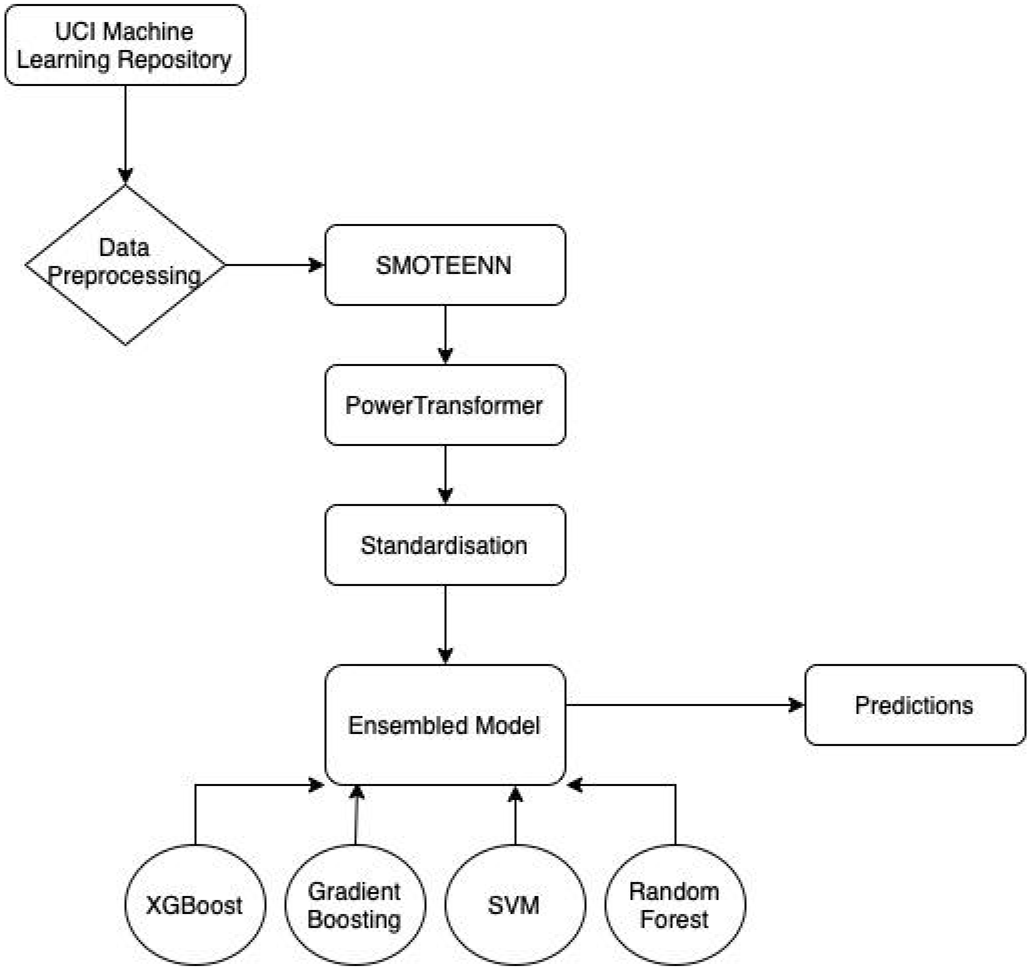


Figure 1 : Proposed Methodology

1.4.1 Preprocessing Phase

The dataset which is used in a form of matrix is highly skewed and contains imbalanced class distribution which can lead to inaccurate prediction hence low accuracy and precision .Also,the matrix contains many duplicate rows which can also contribute to skewed predictions .Initially duplicate rows from the matrix are removed. To address the issue of class imbalance we have used SMOTE-EEN technique which generates synthetic data points and PowerTransformer library to correct the skewness of attributes .After preprocessing ,the dataset is divided into 80% training and 20% testing dataset.

Algorithm 1: Preprocessing Phase

```
Let $S =\{s_1,s_2,s_3,s_4\ldots s_n\}$ be the given dataset

$S_A= D(S)$                    //removing duplicates

$S_B= C(S_A)$                  //using SMOTEEN
$S_C=P(S_B)$                   //using Power Transformer to correct skewness

$S_D= N(S_C)$                  //Normalising dataset

$X=80\%$ training dataset  $X \in S_D$

$Y=20\%$ testing dataset   $Y \in S_D$
```

1.4.2 Training Phase

After removing duplicates ,addressing class imbalance and correcting skewness of the dataset.We can now run different Machine learning algorithms on 80% training dataset and judge the algorithms by taking Accuracy as a metrice.Top performing ML algorithms will be taken into account which can be stacked together .

Algorithm 2: Training Phase

```
Let $R=\{r_1,r_2,r_3\ldots r_n\}$ be different ML algorithms

$E= B(R)$                      //selecting top 4 highest performing algorithm
```

```
L= meta classifier

T= X[E(L)]                              //Stacked model on X
```

1.4.3 Testing Phase

In this phase,Stacked Model parameters are tested on the 20% testing dataset.Initially the Accuracy is tested then classification report is generated .The Accuracy of our proposed algorithm is 98.36% which outperforms the previous literature work .

Algorithm 3: Testing Phase

```
Begin

    P=T(Y)                         //Stacked Model on Y

End
```

Table 4: Symbols used in Algorithm 1,Algorithm 2 and Algorithm 3

S. No.	**Symbols**	**Meaning**
1.	S	Attributes of dataset
2.	D	Remove Duplicate rows
3.	C	SMOTEENN
4.	P	Power Transformer
5.	Z	Normalisation
6.	X	Training set
7.	Y	Testing set
8.	R	ML Models
9.	B	Selecting top 4 models
10.	L	meta classifier
11.	T	Stacked Model on X
12	P	Predictions

Chapter 2

Literature Survey

This section explains the 5 different research work done previously and how they approached the problem and their methodologies.Every year, an estimated 600 million individuals – about one in every ten people in the globe – become unwell after eating contaminated food, and 420 000 die, resulting in the loss of 33 million healthy life years (DALYs)[3].As a result,researchers are coming up with different approaches .Few of them are discussed below.In [5] authors have applied algorithms such as Random Forest , Support Vector Machine(SVM) and Naive Bayes.Along side testing accuracy the author tested training accuracy as well.

S. No.	Author (s)	Approach	Dataset	Performance metrics
1	Cortez et al. [14] 2009	Neural networks,support vector machine,Multiple regression	UCI	Accuracy
2	Appalasamy et al.[16] 2012	Naive Bayes,ID3	UCI	Accuracy
3	Er & Atasoy, [15] 2016	KNN,Support Vector Machine	UCI	Accuracy

4	Gupta[17]2018	Neural Networks,Support Vector Machine	UCI	Accuracy
5	Agrawal et al.[18]2020	Support Vector Machine,Random Forest,Naive Bayes	UCI	Accuracy,Recall, Precision
6	Chao Ye et al.[19]2020	XGBoost,LightGBM	UCI	Accuracy,Recall,Precision
7	Tingwei [25] 2021	KNN,Active learning	UCI	Accuracy
8	Abedin et al.[26]2018	Linear Discriminant Analysis,Multinomial Logistic Regression,Random Forest,Support Vector Machine	UCI (Forina et al., 1991)	Accuracy,Recall,Precision
9	Lee et al.[29]2015	Decision Tree	UCI	Accuracy ,Precision, Recall
10	Wie CC [30] 2012	LAGD Hill Climbing model	UCI	ROC-AUC

Machine Learning algorithms have revolutionised how data analytics and data mining works.Many researchers since the data set was made available had used robust models and different metrics to achieve better results.Cortez et al. 2009 [14] used simple Multiple regression, support vector machine and Neural Networks.On the other hand Er & Atasoy, 2016 [15] used 4 different techniques to experiment with the results but the models remained the same which included Support Vector Machine,Random Forest,k-Nearest Neighbourhood.The first technique was cross-validation,followed by percentage split,cross-validation(after PCA) and percentage split after using PCA.Cross-validation after PCA resulted in highest accuracy among all the methods used.This influenced our research work.(Gupta, 2018) [17] experimented by selecting few features and discarding few based on the correlation among the variables.This resulted in the most robust model.Kanika Agrawal et al, 2020[18] used three model SVM,Naive Bayes and Random Forest while taking all features in account .Ahammed and Abedin 2018 [26] used Linear Discriminant analysis on red and wine wines and got considerably high precision , recall values.Lee et al 2015[29] saw the potential in decision tree as the first bagging method .Wie CC 2012[30] reporting was based on ROC-AUC scores .His study was solely based on decision trees.Our study is an advance in Red wine quality prediction as we have taken into account the skewness and standardisation of data.

2.1. My Contribution:

- Proposed Framework consists of Stacking Based Ensemble learning which adds diversity in the classifier.
- Skewness and Guassian distribution and class imbalance is addressed.
- Hyper Parameter Tuning is used in order to select the best parameter for ML model training.
- The performance of the proposed framework is compared with existing literature on the basis of accuracy, precision, sensitivity, precision and F1 Score

Chapter 3

System Development

This section explains the various machine learning classification methods that are used in the proposed framework. Before the final Ensembling of top performing models, other Classifier models were attempted.10 different classifiers were trained on a training data set. After the initial training 4 models were selected based on their accuracy measure.

A. Random Forest

The random forest classifier is made up of a series of tree classifiers, each of which is constructed using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to categorise an input vector [7].To grow a tree, the random forest classifier utilised in this study uses randomly selected characteristics or a mixture of features at each node. For each feature/feature combination chosen, bagging, a method of generating a training dataset by randomly drawing with replacement N samples, where N is the size of the original training set [8] , was employed. Any examples (pixels) are categorised by selecting the class with the highest number of votes from all tree predictors in the forest [7]

B. K-Nearest Neighbour (KNN)

The k-Nearest-Neighbours (kNN) approach is a basic but effective non-parametric classification method [9]. To classify a data record t, its k nearest neighbours are collected, forming a neighbourhood of t. Majority vote among data records in the neighbourhood is commonly used to determine the classification for t, with or without taking distance-based weighting into account. However, in order to use kNN, we must first select an appropriate value for k, and the classification's success is heavily dependent on this number. In certain ways, the kNN approach is influenced by k.

C. Support Vector Classifier

SVMs are based on statistical learning theory and try to determine the position of decision boundaries that result in the best class separation[10]. SVMs choose the one linear decision boundary that leaves the biggest margin between two classes in a two-class pattern recognition task when classes are linearly separable. The margin is defined as the total of the distances from the nearest points of the two classes to the hyperplane[10]. Using traditional Quadratic Programming (QP) optimization techniques, this problem of maximising the margin can be solved.

Support vector machines (SVMs) [17] are extensively used as a classification tool in a variety of areas. They map the input (x) into a high-dimensional feature space ($z = \phi(x)$) and construct an optimal hyperplane defined by $w \cdot z - b = 0$ to separate examples from the two classes. For SVMs with L1 soft-margin formulation, this is done by solving the primal problem

$$\min \frac{1}{2}||w||^2 + C\sum_i \xi_i$$

$$\text{s.t } y_i(w \cdot z_i - b) \geq 1 - \xi_i ,\ \xi_i \geq 0\ \forall i ,$$

where x_i is the ith example and y_i is the class label value which is either +1 or −1.This problem is computationally solved using the solution of its dual form

$$\max \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$\text{s.t } 0 \leq \alpha_i \leq C\ \forall_i, \quad \sum_i y_i \alpha_i = 0$$

where $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ is the kernel function that performs the non-linear operation.

SVMs were designed for binary classification but can be extended for multiple class classification scenarios that are common. Two main approaches have been suggested for multiclass classification by SVM. In each, the underlying basis has been to reduce the multiclass problem to a set of binary problems, enabling a basic SVM approach to be used. The two approaches are, however, very different in detail. In the “one-against-all” approach, a set of binary classifiers, each trained to separate one class from the rest.

D.Naive Bayes Classifiers [11]

Statistical classifiers are Bayesian classifiers. They are capable of predicting class membership probabilities, such as the likelihood that a given sample belongs to a specific class. The Bayesian classifier is based on the theorem of Bayes. The influence of an attribute value on a particular class is assumed to be independent of the values of the other attributes by Bayesian classifiers. This is known as class conditional independence. It is designed to simplify the computation and is hence termed "naive."

E.XGBoost [24]

XGBoost is a scalable gradient boosting system that focuses on speed and performance. Intelligent tree penalization, proportionate leaf node reduction, and other randomization settings make it apart from traditional gradient boosting algorithms.

F.Ensemble learning [27]

Ensemble learning learns by executing ‘base learner’ multiple times.The final vote is casted on the hypothesis and final weights are executed on ‘meta models'.Various types of Ensemble technique includes Bagging and Boosting

Let D= $\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$be the dataset,where $x_i \in X,y_i \in Y$= $\{c_1,c_2, ...,c_l\}$ and N is the total number of instances in the dataset and c_1 , c_2 , ... , c_l are the class labels. Typically the input space X consists of many features and hence its elements are represented as N-tuples of ‘d’ dimensions.

$$X = \{(x_{11}, x_{12}, \cdots,x_{1d}),(x_{21}, x_{22}, \cdots,x_{2d}), \cdots,(x_{N1}, x_{N2}, \cdots,x_{Nd})\}. \quad (1)$$

where ‘d’ is the number of features in the input space. Hence, the dataset is represented as:

$$D= \{((x_{11}, x_{12}, \cdots,x_{1d}),y_1), ((x_{21}, x_{22}, \cdots,x_{2d}), y_2), \cdots,((x_{N1}, x_{N2}, \cdots,x_{Nd}),y_N)\}. \quad (2)$$

Split D into two sets S_1 and S_2 at 80% and 20% respectively. S_1 is used as a training set and S_2 is used as a test set.

$$D = S_1 \cup S_2. \quad (3)$$

Base Tier

Let $WC = \{WC_{1,} WC_2, \dots , WC_m\}$ be the set of weak classifiers.

Let T_1 , T_2 , $\cdots$, T_{10} be the equal size sets, partitioned from the set S_1 .

Fold 1: The Training set $T = T_1 \cup T_2 \cup ... T_9$.

The testing set $V = T_{10}$

Train the weak classifiers using T.

The weak classifiers WC_m maps the elements of T into one of the classes in Y.

$$WC_k{:}X_g \rightarrow Y \tag{4}$$

$$WC_k{:}X_g \rightarrow \{c_1,c_2,\dots,c_l\} \tag{5}$$

where k=1,2,...,m and X_g belongs to T i.e.$X_g \in T$ (6)

In the next fold, *T*9 is taken as the testing set and the remaining nine sets combined together are taken as the training set. This is repeated until all the partitioned sets in S_1 are taken as testing sets. This results in the predictions for all the instances in T.Let $WL = \{WL_1, WL_2, \dots , WL_m\}$ be the set of labels predicted by the weak classifiers

i.e.$WL_k = WC_k(X_g)$ (7)

Where k = 1,2….m.

Ensemble Tier

Let EC= $\{EC_1, EC_2, \dots,EC_m\}$be the set of combination schemes and they combine the label predicted by the weak classifiers and map them into one of the classes in Y.

$$EC_k{:}WL_k \rightarrow Y \tag{8}$$

EC_k:WL_k

$$EC_k\text{:}WL_k \rightarrow \{c_1, c_2, \ldots, c_l\}. \tag{9}$$

$$EC_k = Aggregation\ of\ (WL_k)$$
(10)

$$EC_k = Aggregation\ of\ (WC_k\ (X_g)).\ \text{where k = 1, 2, ..., m} \tag{11}$$

Let EL= $\{Y_{ec1}, Y_{ec2}, \cdots, Y_{ecm}\}$be the set of labels predicted by the ensemble classifiers.This can be represented in mathematical form as

$$Y_{eck} = EC_k(WL_k), \tag{12}$$

$$Y_{eck} = EC_k\ (WC_k(X_g)).$$
(13)

Where k = 1,2…m.

G.SMOTEENN [32]

This technique is very helpful in solving class imbalance as it generates synthetic data points by SMOTE using the ENN algorithm.Synthetic data points are very different from duplicate points.

F.Gradient Boosting Algorithm [31]

The gradient boosting approach may predict both continuous and categorical target variables . Mean Square Error (MSE) is the cost function when used as a regressor, and Log loss is the cost function when used as a classifier.

3.1 Proposed Framework

In this section we explained model selection criteria and parameter setting of different algorithms used in the building the framework. Experimental setup & proposed methodology has been explained further.

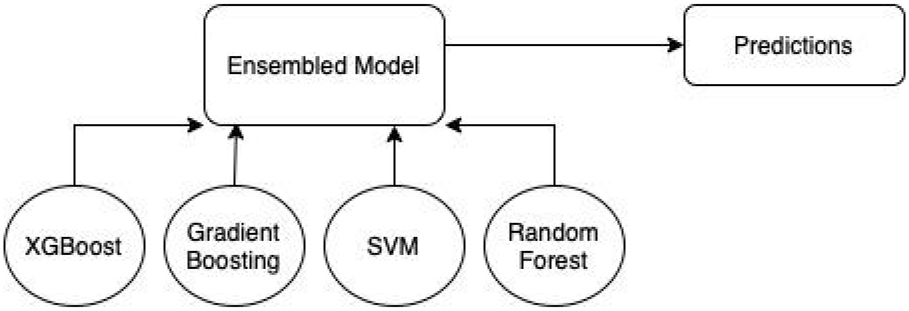


Figure 2: The proposed ensemble framework for wine quality prediction

3.2 Model Selection

Figure 1 all-inclusive depits proposed framework for red wine quality prediction. Firstly, we have taken Data from UCI Machine Learning repository(only red wine data),which is explained in Section 3.3.We deleted outliers after thoroughly analysing the data and discovering correlations among other parameters.. Then split our data into two partitions: Training Data consisting of 80% of instances & Testing Data with 20% instances.As Chao Ye et al 2020[19] used XGBoost and LightGBM gave highest accuracy we included these models after considering every predefined models and these models helped to increase the overall accuracy. Considering Red wine quality prediction literature review most of the authors who achieved considerable accuracy Agrawal et al. 2020[18] , Cortez et al. 2009[14],Er & Atasoy 2016[15] ,Gupta, 2018[17] have used SVM .Naturally SVM was to be included in model selection section. Various Bagging and Boosting algorithms are proven to increase the accuracy considerably as we can see in Chao Ye et a.l 2020[19] thats why apart from LightGBM and XGBoost other algorithms including Gradient Boosting algorithm ,Decision Tree,Random Forest were considered while selecting models for stacked ensemble based classifier.

Apart from selecting our models using Literature survey ,Hyperparameter tuning is also done to maximise the accuracy which is discussed in Section 4.2 .After our model was built , We used Stacked Classifier class to perform ensemble learning that uses meta classifiers on specified base learners chosen by us.The aim is to have a diverse set of learners together .Given various classifiers we choose the one with higher accuracy to be used as one of the base learners .

3.3 Parameter Setting

In this section, we discussed various factors that were used to improve the accuracy of our Stacked Ensembled Model.As XGBClassifier,Random forest,SVM,Gradient Boosting Classifier were with highest accuracy ,These four were chosen as a base learner for the ensemble model. Hyperparameter tuning on the following model was done to further improve the accuracy which as shown in Table 2 ,can in turn improve the accuracy of stacked classifiers. Random State is 42 throughout.

Table 2 : Hyperparameters used

Model	Learning Rate	N estimators	C Value	Kernel	Gamma
Gradient Boosting	1	300	-	-	-
XGBoost	0.1	500	-	-	-
SVM	-	-	10	rbf	1

3.4 Experimental Setup

3.4.1 Data set

The Data was retrieved from UCI Machine Learning Repository[12-13].It contains 11 input variables based on physicochemical tests which includes fixed acidity,volatile acidity,citric acid,residual sugar,chlorides,free sulfur dioxide,total sulfur dioxide,density,pH,sulfates,alcohol.The output variable called quality contains variable indicating lowest quality wine with 3 going upto 8 which depicts good quality wine.This makes it a multiclass classification problem. Table 3 gives a detailed description of the attributes given in the dataset.

Table 3: Description of Nominal Attributes

Attribute	Description
Fixed acidity	Can be fixed or non-volatile 4.6 (least acidic) to 15.9(Most acidic)
volatile acidity	adds to the unlikable taste 0.12 to 1.58
Citric acid	increases the originality of wine 0 to 1
Residual Sugar	Most wines have dense accumulation of sugar 0.9 to 15.5
Chlorides	the salt content of the wine 0.01 to 0.61

Free sulphur dioxide	Helps shutting of fermentation 1 to 72
total sulphur dioxide	healthy in low amount but disagreeable if found high high concentration 6 to 289
Density	It hangs on the quality of water 0.99 to 1
pH	Abundantly wines have 3-4 pH 2.74 to 4.01
Sulphates	Helps preventing uneasiness 0.33 to 2
Alcohol	the wine's alcohol content in percentage 8.4 to 14.9
Quality	variable output (based on sensory data, score between 3 and 8) Class 1: 3 Class 2: 4 Class 3: 5 Class 4: 6 Class 5: 7 Class 6: 8

First of all,class imbalance is analysed in the dataset, as it as multiclass classification problem and we are predicting values for each class it is essential to address the imbalance by

SMOTEEN.Furthur the few features were highly skewed which could have resulted in partial results as a result columns whose skewness was greater than 0.75 were corrected by using PowerTranfomer library .Other than quality which contains discrete values every other column contains continuous values.

3.4.2 Data visualisation

Multicollinearity can drastically affect the prediction of a Machine Learning algorithm. Multicollinearity affects the precision of computed coefficients, lowering your regression model's statistical power. You might not be able to rely on p-values to detect statistically significant independent variables[28].Figure 2, shows Multicollinearity between different attributes.As we can see , this data set is free from multicollinearity but there is a heavy class imbalance as shown in Fig 4, we have corrected the imbalance by using SMOTE-ENN library.As shown in Figure 3 few attributes are highly skewed which can make the prediction wrong , to correct this we identified highly skewed columns (whose skewness was greater by 0.75) and were corrected by using PowerTransformer library .Another example of class imbalance can be seen in Figure 5 where we can thoroughly examine the distribution of class among various attributes

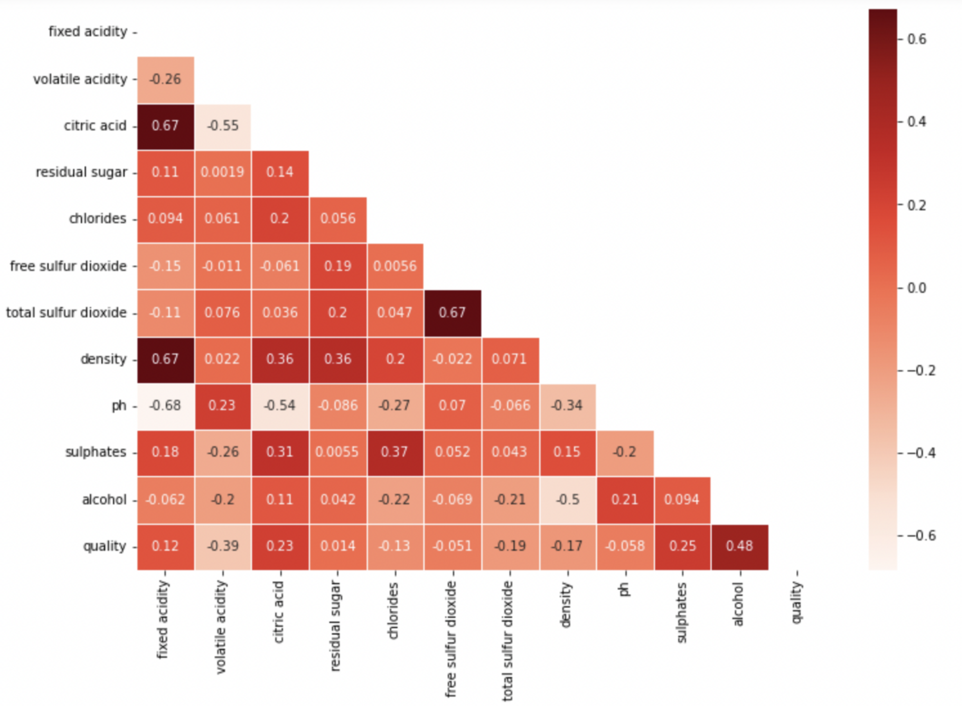
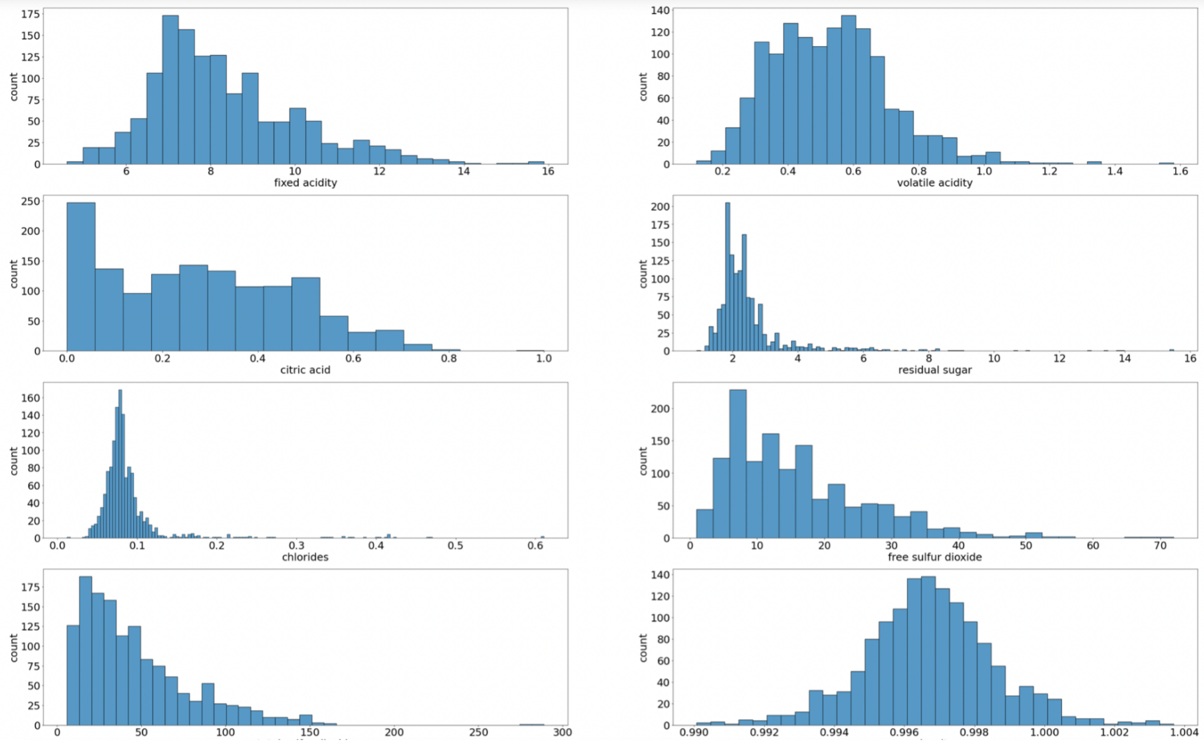


Figure 2 : Multicollinearity of attributes



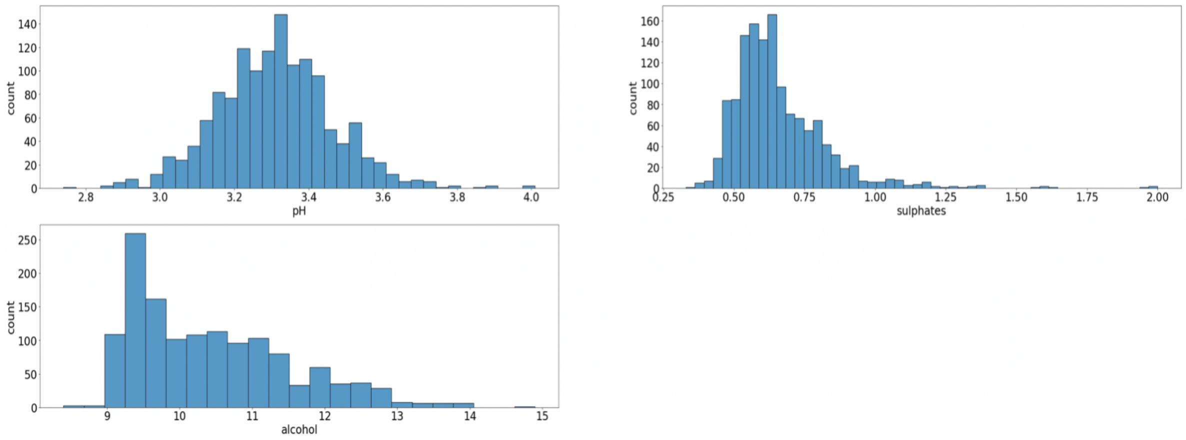


Figure 3 : Skewered attributes (X-axis: Total occurrence of an attribute Y-axis:attribute)

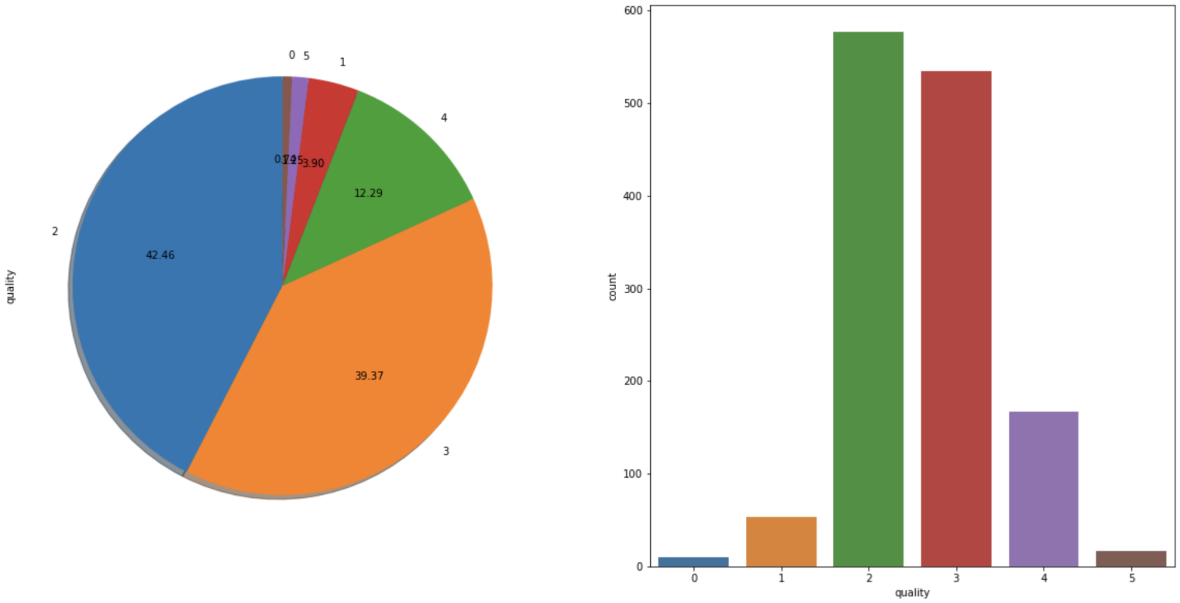
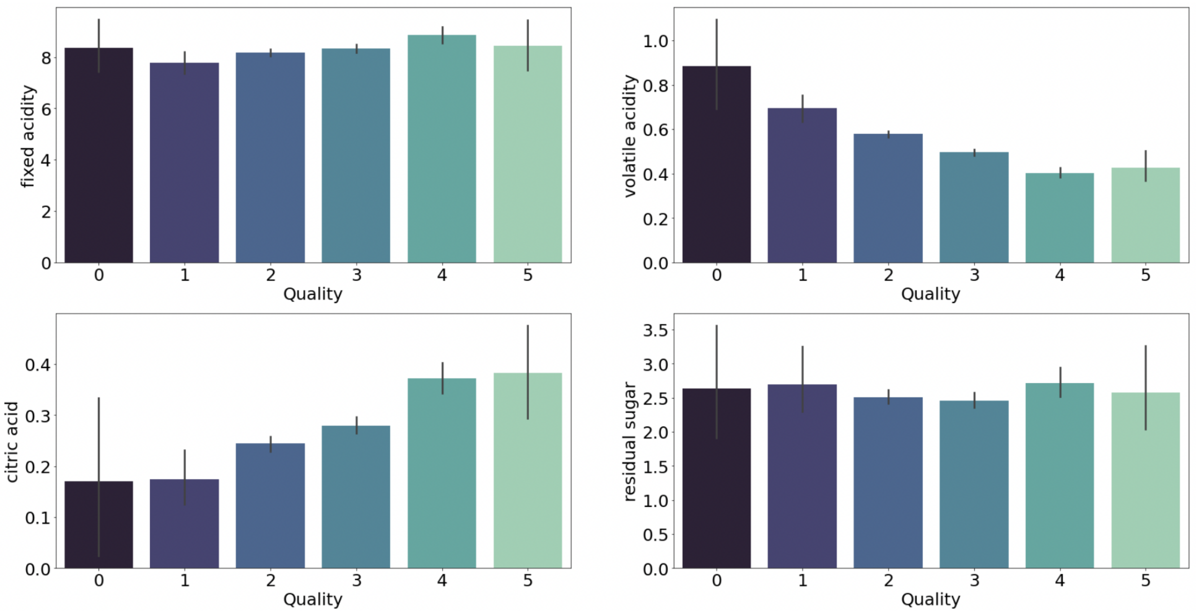
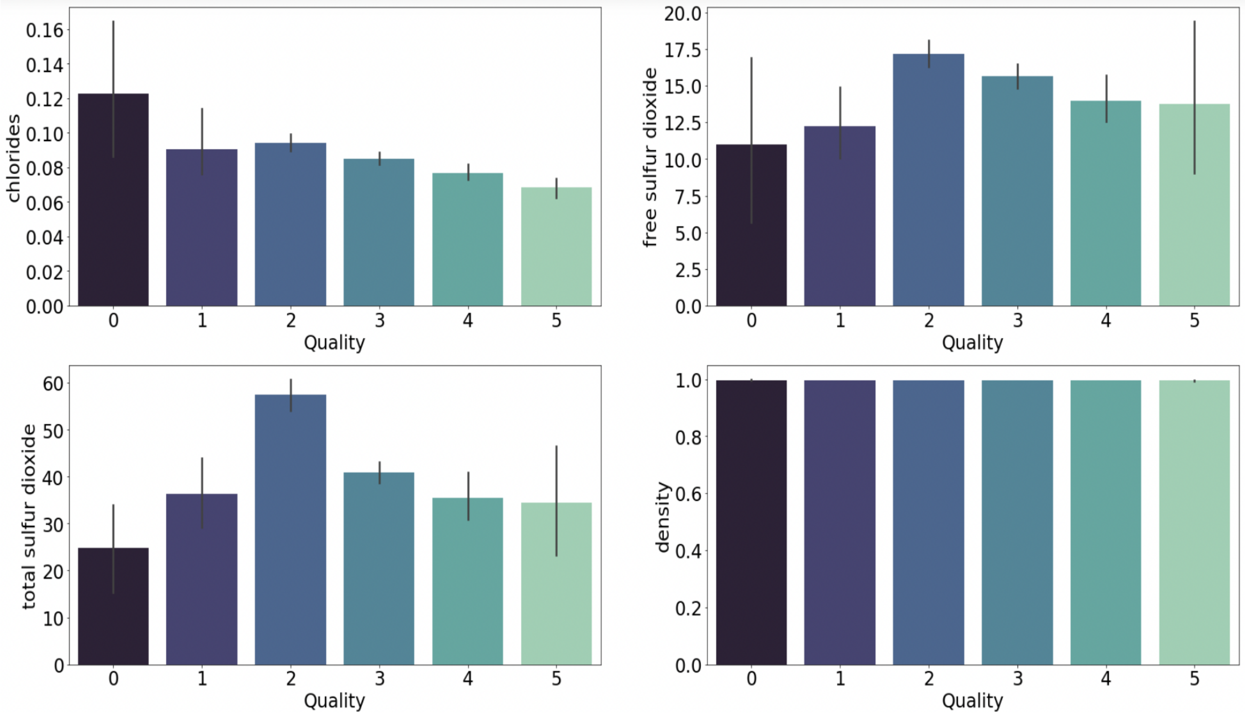


Figure 4 : Class Imbalance(Pie Chart represents distribution of each class, Bar Chart- X-axis : Total count of target variable Y-axis: Target variable)


fixed acidity
0
2
4
6
8
Quality
0 1 2 3 4 5
volatile acidity
0.0
0.2
0.4
0.6
0.8
1.0
Quality
0 1 2 3 4 5
citric acid
0.0
0.1
0.2
0.3
0.4
Quality
0 1 2 3 4 5
residual sugar
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Quality
0 1 2 3 4 5


chlorides
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Quality
0 1 2 3 4 5
free sulfur dioxide
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Quality
0 1 2 3 4 5
total sulfur dioxide
0
10
20
30
40
50
60
Quality
0 1 2 3 4 5
density
0.0
0.2
0.4
0.6
0.8
1.0
Quality
0 1 2 3 4 5

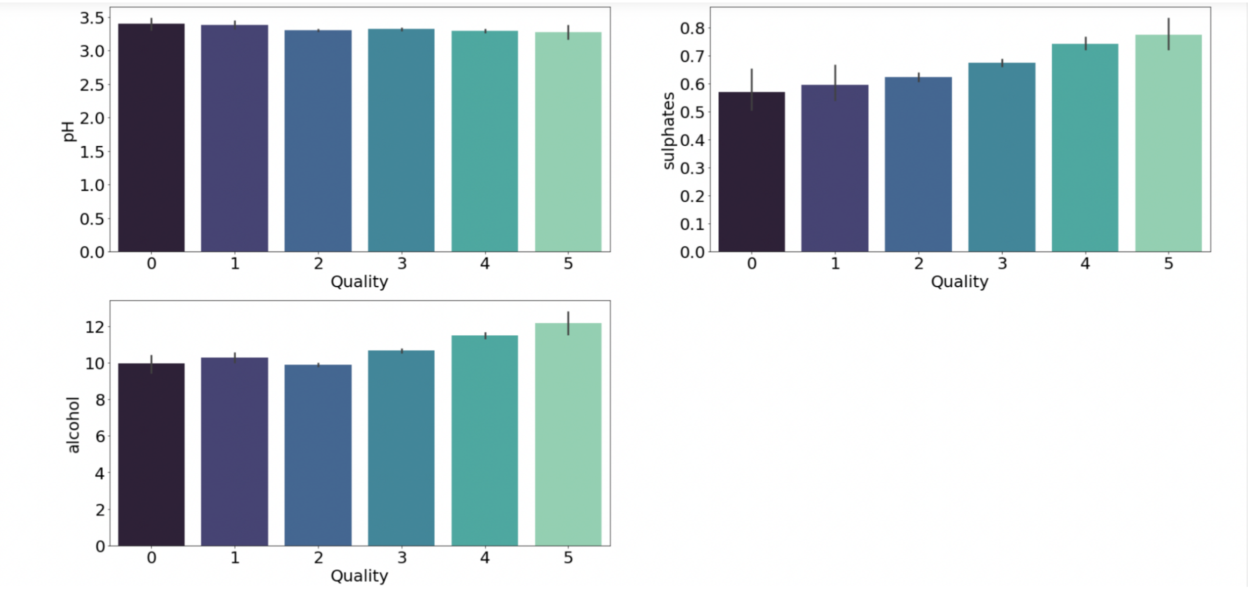


Figure 5 : Analyzing each column using barplot with quality

Chapter 4

Performance Analysis

In this section we have discussed the results & analysis of our proposed framework. Different performance metrics have been used to evaluate the algorithms. Further we have compared our model with other existing models and its comparison with respect to accuracy, precision, sensitivity, precision, F1 Score, ROC & MCC. We have also discussed proposed models with different algorithms and models covered in Section 2.

4.1 Performance Metrics

The following five parameters are used to assess the performance of the proposed framework:

1.Accuracy: The value predicted when the sum of True Positive and True Negative is divided by the sum of True Positive, False positive, False Negative and True Negative values of a confusion matrix.

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + False\ Negative + True\ Negative)}$$

2.Precision: The value obtained when True Positive is divided by the sum of True Positive and False Positive values of a confusion matrix

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

3.Recall: Sensitivity sometimes also known as Recall. It is the value obtained when True Positive is divided by the sum of True Positive and False Negative values of a confusion matrix.

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

5.F-Measure: F1 Score is obtained by multiplying Recall and Precision divided by sum of Recall and precision of a confusion matrix. Result is then multiplied by two.

$$F1\ Score = \frac{2 * (Recall * Precision)}{(Recall + Precision}$$

4.2 Comparison with ML Models

We created several baseline models but the models with highest accuracy were chosen for stacking purposes as ensemble modelling adds more diversity to the predictions.We have chosen accuracy as the metrics to judge models initially.Accuracy of all the models used is shown in Table 5.Top 4 best performing models which includes XGBClassifier,Random forest,SVM,Gradient Boosting Classifier are stacked together to give even better accuracy.We also tested out proposed algorithm on various factors including Accuracy,precision,recall,F1-Score.Figure 7 graphically represent the comparison of different Machine Learning algorithms with our proposed algorithm.As we can see our proposed algorithm outperforms the existing algorithm and previous literature work as shown in Figure 8.Hence,our work is an advance is Red wine classification.We stacked these classifiers to get accuracy of 98.36% .As it is a multiclass classification problem we got an average of 98.0% precision and 98% recall. As shown in Table 6 besides accuracy we have calculated Precision, Recall and F1 score for comparison with other algorithms.

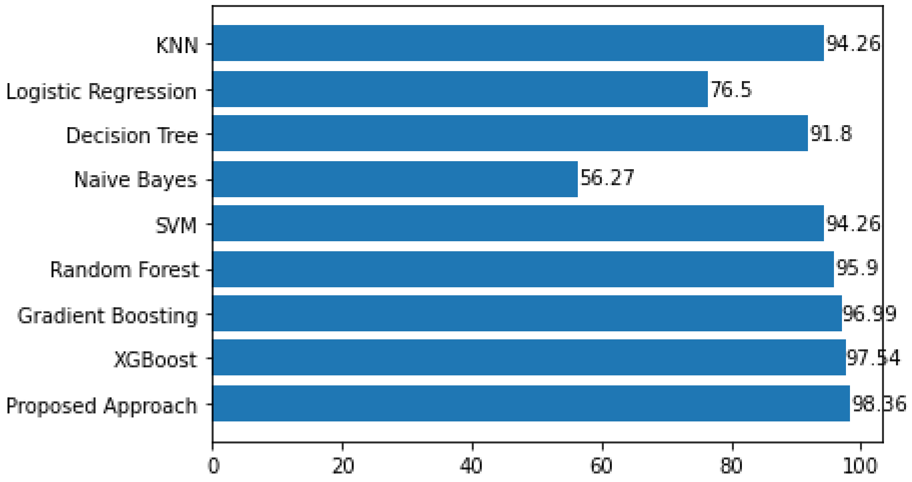
Additionally,to analyze our models we have built ROC (receiver operating characteristic curve) because it shows the tradeoff between specificity and sensitivity for every combination of tests.As we can see in Figure 9 our proposed algorithm ROC Curve is approximately perfect.The better the model the closer the area of ROC is to 1.As our problem is MultiClass Classification ROC curve has taken macro average into account . Figure 8 depicts Multi-Class Classification more comprehensively by plotting the curve for each class through our proposed framework .Ensembled model helped us to add diversity and multiplicity in our model. Further stacked based models add assortment which means if an individual model gives a wrong prediction about a certain feature ,another model used in a stacked based ensemble may have a chance to correctly identify the same feature .

Our work greatly contributes towards food/wine analytics as we are able to classify good and worst quality of wine while outperforming the literature review.This can greatly impact the future research work can can nearly perfectly predict the correct quality of the food item.

Table 5: Comparison of ML algorithm & their respective accuracies

S. No.	**Algorithm**	**Accuracy**
1	XGBoost Classifier	**97.54%**
2	**Gradient Boosting Classifier**	**96.99%**
3	**Random Forest**	**95.90%**
4.	**SVM**	**94.26%**
5	KNeighbors Classifier	94.26%
6	DecisionTree Classifier	91.80%
7	Logistic Regression	76.50%
8	Naive Bayes	56.27%

Figure 7: Comparison of accuracy of proposed framework with different ML models



Accuracy

Table 6: Comparison of proposed Framework with existing ML Models

Models	Precision	Recall	F1	Accuracy
Random Forest	0.96	0.96	0.95	95.90%
Gradient Boosting	0.97	0.97	0.96	96.72%
XGBoost	0.98	0.98	0.97	97.54 %
Logistic Regression	0.75	0.77	0.75	76.50%
KNN	0.94	0.94	0.94	94.26%
SVM	0.94	0.94	0.94	94.26%
Decision Tree	0.92	0.92	0.92	92.34%
Chao Ye et al 2020[19]	88.15	88.67	88.41	91.04%
Proposed Methodology	0.98	0.98	0.98	98.36%

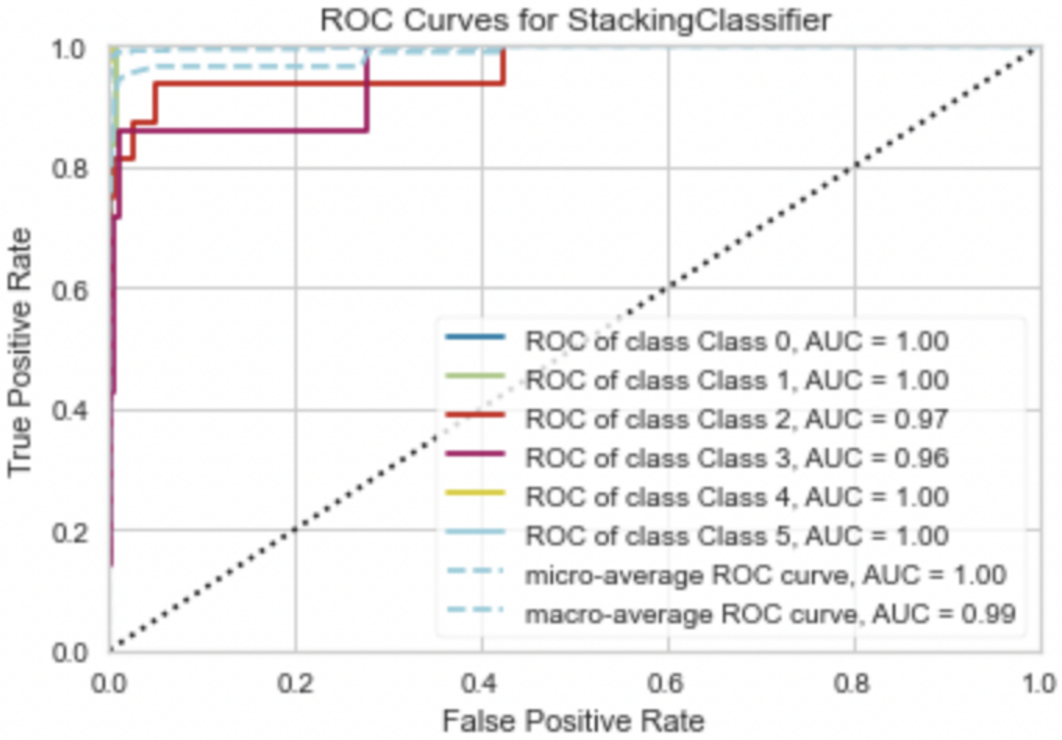


Figure 8 : ROC-AUC Comparison for each class

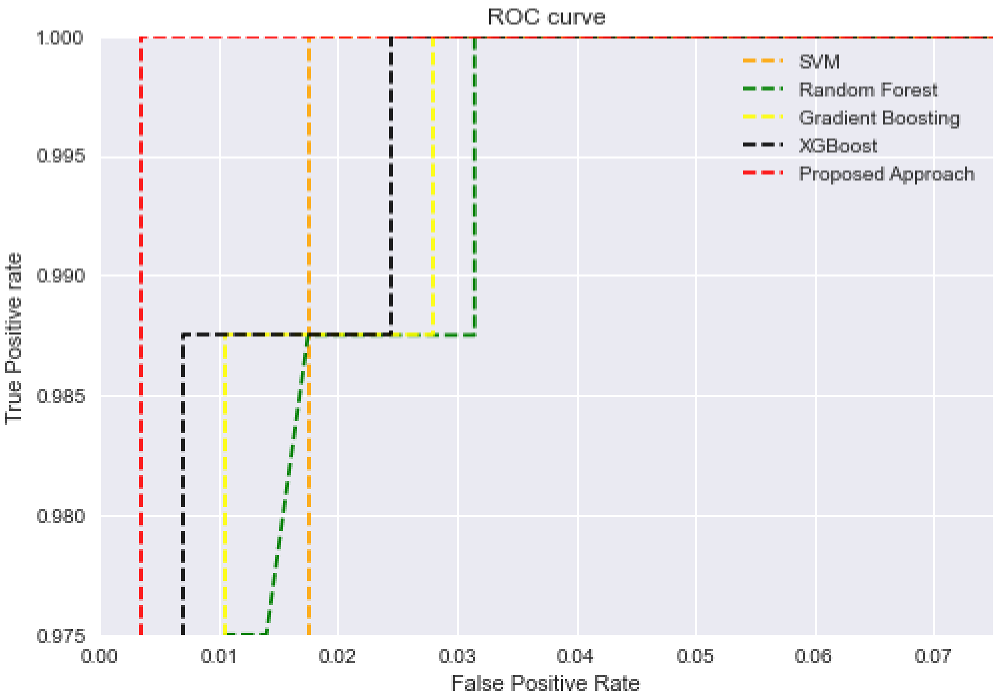


Fig 9 : ROC Curve comparison

4.3 Comparison with Existing Literature

Our proposed algorithm shows a perfect ROC curve and good accuracy and it can be considered an advance in Red wine quality prediction in return an advance in classifying quality of any other food item. Chao Ye et al. [19] used XGBoost which influenced our work ,Further every author used SVM,Random Forest which heavily influenced our work As shown in Figure 10 our proposed methodology outperforms previous literature work done on the dataset . This can be used further in biomedical research work relating to food/water quality predictions.Applying Ensembled based Stacking can benefit further research work as it provides diversity to classifiers and increases other parameters like Accuracy , Precision etc.

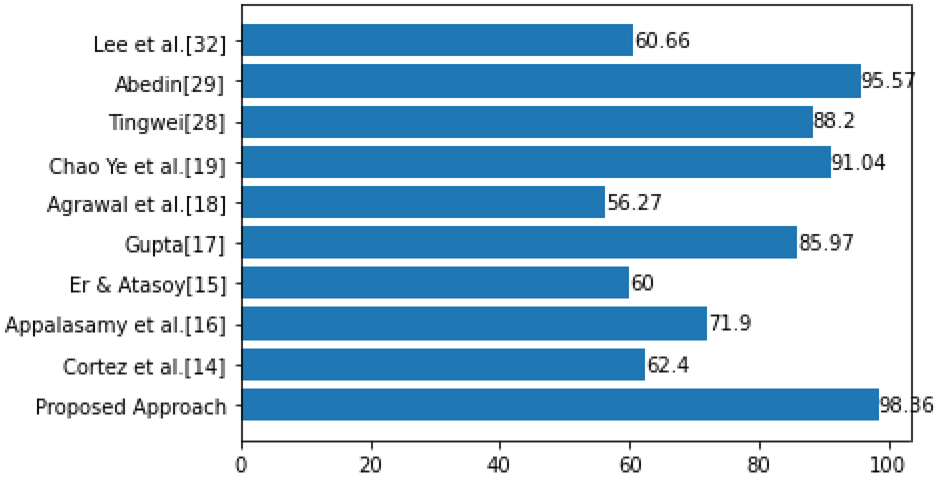


Fig 8: Comparing proposed model with existing literature

Chapter 5

Conclusions

In this paper, we offer a machine learning-based computational framework for predicting Red Wine Quality. The proposed framework successfully sorted red wines into various classifications. The proposed framework's key contribution is handling skewed and imbalanced data utilising a power transformer and the SMOTE-ENN technique. Furthermore, ensemble learning helps to increase variety among the base learners, which increases prediction accuracy. Precision, Specificity, Recall, and F1 Score are used to evaluate the performance of all approaches.On a single benchmark dataset, all algorithms are trained and tested. Most present strategies do not take into account the uneven nature of data and skewed data when creating Red Wine Quality prediction tools.The proposed framework addresses the challenges of class imbalance and data skewness.

References

[1] Renee J. Consequences of Poor Quality in Food. Healthfully. Published March 7, 2010. Accessed April 12, 2022. https://healthfully.com/90778-consequences-poor-quality.html

[2]*Diet, Nutrition, and the Prevention of Chronic Diseases : Report of a Joint WHO/FAO Expert Consultation [Geneva, 28 January - 1 February 2002].* World Health Organization; 2003. Accessed May 12, 2022. https://www.who.int/publications/i/item/924120916X

[3] World. Food safety. Who.int. Published April 30, 2020. Accessed May 12, 2022. https://www.who.int/news-room/fact-sheets/detail/food-safety

[4] Andersen A, Due P, Holstein BE, Iversen L. Tracking drinking behaviour from age 15–19 years. Addiction. 2003 Nov;98(11):1505-11.

[5] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104095.

[6] Kumar R, Pal R. India achieves WHO recommended doctor population ratio: A call for paradigm shift in public health discourse! J Family Med Prim Care. 2018;7(5):841-844. doi:10.4103/jfmpc.jfmpc_218_18

[7] Breiman L (1999) Random forests—random features. Technical Report 567, Statistics Department. University of California, Berkeley. ftp://ftp.stat.berkeley.edu/pub/users/breiman

[8] Breiman L. Bagging predictors. Machine learning. 1996 Aug;24(2):123-40.

[9] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[10] Vapnik V. The nature of statistical learning theo~.

[11] Leung KM. Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering. 2007 Nov;2007:123-56.

[12] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties
[13] In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[14] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. “Modeling wine preferences by data mining from physicochemical properties.” In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[15]Er Y, Atasoy A. The classification of white wine and red wine according to their physicochemical qualities. International Journal of Intelligent Systems and Applications in Engineering. 2016 Sep 1;4(Special Issue-1):23-6.

[16]Appalasamy P, Mustapha A, Rizal ND, Johari F, Mansor AF. Classification-based data mining approach for quality control in wine production. Journal of Applied Sciences. 2012 Jun;12(6):598-601.

[17]Gupta Y. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science. 2018 Jan 1;125:305-12.

[18]Kumar S, Agrawal K, Mandan N. Red wine quality prediction using machine learning techniques. In2020 International Conference on Computer Communication and Informatics (ICCCI) 2020 Jan 22 (pp. 1-6). IEEE.

[19]Ye C, Li K, Jia GZ. A new red wine prediction framework using machine learning. InJournal of Physics: Conference Series 2020 Nov 1 (Vol. 1684, No. 1, p. 012067). IOP Publishing.

[20]Karlsson B. Wine Consumption In The World 2020 In Decline, A Detailed Look. *Forbes*. https://www.forbes.com/sites/karlsson/2021/12/31/wine-consumption-in-the-world-2020-in-decline-a-detailed-look/?sh=3c0dcc0e3f71. Published January 16, 2022. Accessed April 9, 2022.

[21]Ritchie H, Roser M. Alcohol Consumption. Our World in Data. Published April 16, 2018. Accessed April 10, 2022. https://ourworldindata.org/alcohol-consumption

[22]Golan, R., Gepner, Y. & Shai, I. Wine and Health–New Evidence. *Eur J Clin Nutr* 72, 55–59 (2019). https://doi.org/10.1038/s41430-018-0309-5

[23]Abu-Amero KK, Kondkar AA, Chalam KV. Resveratrol and Ophthalmic Diseases. *Nutrients*. 2016;8(4):200. Published 2016 Apr 5. doi:10.3390/nu8040200

[24]Smith J. Is red wine good for you? Medicalnewstoday.com. Published April 21, 2020. Accessed April 11, 2022. https://www.medicalnewstoday.com/articles/265635#can-wine-improve-health

[25]Tingwei Z. Red wine quality prediction through active learning. InJournal of Physics: Conference Series 2021 Jul 1 (Vol. 1966, No. 1, p. 012021). IOP Publishing.

[26]Ahammed B, Abedin M. Predicting wine types with different classification techniques. Model Assisted Statistics and Applications. 2018 Jan 1;13(1):85-93.

[27]Lappalainen, H. and Miskin, J.W., 2000. Ensemble learning. In *Advances in Independent Component Analysis* (pp. 75-92). Springer, London.

[28]Frost J. Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Statistics By Jim. Published April 2, 2017. Accessed April 12, 2022. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[29]Lee S, Park J, Kang K. Assessing wine quality using a decision tree. In2015 IEEE International Symposium on Systems Engineering (ISSE) 2015 Sep 28 (pp. 176-178). IEEE.

[30]Wei CC. Receiver Operating Characteristic for Diagnosis of Wine Quality by Bayesian Network Classifiers. AMR 2012;591–593:1168–73. https://doi.org/10.4028/www.scientific.net/amr.591-593.1168.

[31]Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

[32]Prati, R.C., Batista, G.E. and Monard, M.C., 2004, September. Learning with class skews and small disjuncts. In *Brazilian Symposium on Artificial Intelligence* (pp. 296-306). Springer, Berlin, Heidelberg.